Eemi Kärkkäinen 909512

Task A: Review a related paper

Primary goal of the paper:

The primary goal of the paper is to introduce the reader to research which is done to find out whether biographies on Wikipedia are affected by an effect called glass ceiling effect. This means that researchers tried to find out if females presented in Wikipedia are more notable and must get more recognition in order to get the same visibility as the equivalent male counterpart on average.

As the paper mentioned, there is a subtle glass ceiling effect happening in the structure of Wikipedia's biographies in six of the most popular language editions. The paper's purpose is also to spread the information and findings about gender bias that appears online and influence Wikipedia's guidelines and editors to be more gender neutral.

Main methods:

As there is told in the paper, the research was done in three different dimensions:

1. *Global notability of people according to external and internal proxy measures.*
2. *Topical focus and linguistic bias of biography articles.*
3. *Structural properties of the articles, including meta-data end network-theoretic position of people in the Wikipedia article link network.*

In the first-dimension researchers simply analyzed the different language editions and counted how many languages an individual appears in. They also analyzed Google's search data to see the search volume of an individual.

In the second-dimension researchers used the dictionary method to discover that complex words (such as adjectives) are more likely to have a negative meaning in female biographies than in male and to have more positive meaning in male than female biographies. Researchers also used normalization and term frequency methods to analyze the topical focus in addition to n-grams to find repeating phares or bags of words.

In the third-dimension researchers constructed hyperlink networks to analyze the connectedness of the biographies.

Key findings:

Probably the most important key finding the paper presented was the manifestation of gender bias and the glass ceiling effect. There were many reasons for these effects to surface but the researchers for example figured out that the better access to information has increased the chance that an accomplished female is depicted in Wikipedia. The absence of some information, especially from prior 1900s, makes a significant difference. One important key finding is that the Wikipedia's gender bias cannot only be from Wikipedia mirroring off-line word as the paper states: *"Our empirical results uncover significant gender differences at various levels that cannot only be attributed to the fact that Wikipedia is mirroring the off-line world and its biases."*

Description of a follow-up analysis:

A follow-up analysis based on the paper that could have been done is to examine the connection between female depicted in Wikipedia and their political beliefs. It would be interesting to see if there is a connection. This could have been achieved by counting the individuals straightly based on the declared political party they associated themselves in and analyzing the rest of the biographies in absence of this information.

After the analysis, the conclusions could be made based on the fraction of the female individuals in different political parties. If the resulting fractions were significant, the results could be analyzed further to see how different time periods affected the documentation of otherwise successful individuals.

Task D: Additional paper reviews

Part 1:

Critical examination of the analysis of the dataset:

The dataset used by the researchers isn't from a specific time so one could consider it "always on". Still, one could argue that there is a lot more data from, for example, 1900s and 2000s because documentation of individuals has increased significantly compared to 1800s and prior. The dataset, however, is "nonreactive" because the study isn't affecting the data. All the analyzed data is created prior to the research and the gender data, for example, would be unlikely to change even if the people knew they were participating in a study.

As all "big data" datasets, this dataset is incomplete as it was mentioned earlier. However, it is accessible to anyone and fairly clean, which makes it easy for anyone to do some follow-up research in addition to the research presented by the paper.

In addition, the researchers discussed in the paper that their results are limited to English Wikipedia and biased towards western culture. This was mostly because English Wikipedia has the most biographies.

Part 2:

Primary goal of the paper:

The primary goal of the paper is to examine how conflicts appear in Wikipedia. The research shows that it is possible to develop methods for predicting conflicts from simple metrics. In addition, it also shows how the network of reverts in some Wikipedia articles can be visualized to analyze the causes of the conflict.

The paper brings out some details that seem to cause and some that seem to mitigate the conflicts. These discoveries are presented in the paper to raise awareness on how platforms, such as Wikipedia, could be refined and how, for example, conflicts could be avoided more effectively.

Main methods:

Firstly the conflicts were analyzed in the paper by a bottom-up data driven method. Researchers computed a unique identifier for each article revision with MD5 hashing scheme. This allowed researchers to see when a revision was a revert and thus when the conflicts occurred in Wikipedia. Secondly a top-down user driven method was used to capture the partial reverts by the tags created by users.

Researchers also created their own approaches to analyze and detect conflicts. For example, the paper tells how researchers trained their machine learning model to predict Controversial Revision Count (CRC) numbers purely based on the page statistics. This made detecting the controversial articles automatic and efficient.

In addition, a user conflict model was created by the researchers. This model made identifying user clusters, and thus visualizing the groups involved in the conflicts, possible.

Key findings:

The first key finding was that Wikipedia is still growing exponentially but the number of created articles and the amount content is decreasing. This means that most of the growth goes into the levels of maintenance and indirect work. Furthermore, this indicates that the quantity of

conflicts in Wikipedia has also increased. The researchers also found that the coordination of users and procedures for dealing with conflicts are vital for platforms such as Wikipedia.

Second finding presented in the paper was that it is possible to develop an automatic detection and prediction models of complex phenomena. For example, they found that it is possible to detect controversies in Wikipedia automatically.

Third important finding arose when the researchers were analyzing the level of conflict of an article with machine learner they had trained. They found that an effective way to resolve a conflict is to increase the number of users involved in editing the article. The researchers also discovered that their Revert Graph model could help identifying high-conflict users from such networks constructed by unique editors in Wikipedia articles.

Description of a follow-up analysis:

In a follow-up analysis based on the paper, one could, for example, analyze the length of conflicts. Implementation of this analysis would be straightforward. First, all the article revision should be hashed the same way the paper presented. After that all the articles where conflicts occur could be analyzed and the conclusions could be made. However, one would have to decide the definition of a start for the conflict and the end. Start could be the first revert or the revision which caused the first revert. One could define the end to be the first revision after the last revert. In addition, to see the whole picture, another follow-up analysis could be done to see when the conflict resolved by examining the number of revisions in different states during the conflict.

Critical examination of the analysis of the dataset:

The dataset used by the researchers could be considered “always on.” The data consists of any action made in Wikipedia from the whole life span of the platform so the datapoints have been collected continuously. Dataset is “nonreactive” if the people and their actions included in the data are not affected by the research. The dataset used in the research presented by the paper is nonreactive because the people using the platform are not aware of that they are studied.

The dataset is representative because, for example, some revisions have revision tags and there are also other ways to analyze the characteristics of the conflicts. I think there isn’t much insensitive data in the dataset. All the data gathered to the dataset is public and it can be studied by anyone. Of course, discussions on the platform and some opinions presented could tell something from a person but couldn’t make it possible to recognize them.

The paper discusses that there are some limitations to the dataset. For example, vandalism, inaccuracies and other quality issues might create noise to the results of the research.